

# An efficient algorithm for the entropy rate of a hidden Markov model with unambiguous symbols

Jaideep Mulherkar  
Dhirubhai Ambani Institute of  
Information and Communication Technology  
jaideep.mulherkar@daiict.ac.in

February 19, 2014

## Abstract

We demonstrate an efficient formula to compute the entropy rate  $H(\mu)$  of a hidden Markov process with  $q$  output symbols where at least one symbol is unambiguously received. Using an approximation to  $H(\mu)$  to the first  $N$  terms we give a  $O(Nq^3)$  algorithm to compute the entropy rate of the hidden Markov model. We use the algorithm to estimate the entropy rate when the parameters of the hidden Markov model are unknown. In the case of  $q = 2$  the process is the output of the Z-channel and we use this fact to give bounds on the capacity of the Gilbert channel.

**Keywords:** Entropy rate, Hidden Markov model, Algebraic measures, Gilbert channel capacity

## 1 Introduction

Entropy rate of a stationary stochastic process  $\{X_n\}_{n=0}^\infty$  is the limit

$$H(\mu) = \lim_{n \rightarrow \infty} \frac{S_n(X_1, X_2, \dots, X_n)}{n} \quad (1)$$

where  $\mu$  is the measure associated with the process and  $S_n(X_1, X_2, \dots, X_n)$  is the joint entropy of  $\{X_1, X_2, \dots, X_n\}$ . It amounts to the average amount of information per symbol. In this paper we study the entropy rate of a hidden Markov process (HMP) that has at least one unambiguous received symbol. A received symbol is unambiguous if after receiving the symbol one can conclude with certainty the state or input symbol. An example of an HMP with one unambiguous received symbol is the output of the Z-channel with Markov input process which has been used to model optical communication systems. A closed form formula exists when the process is Markov however a tractable formula for the entropy rate of a general HMP is still an outstanding problem. Entropy rate of a HMP was first studied by Blackwell in 1957 [1]. Blackwell showed

that the entropy rate of a HMP can be computed as an integral of a function defined on the simplex with respect to a measure. Unfortunately in most cases the measure is quite complicated and computation of the entropy rate using this method is not feasible. Birch [2] showed that the entropy rate can be upper and lower bounded by functions that converge exponentially fast to the entropy rate. A formula for the entropy rate of a HMP also assumes importance because of the use of hidden Markov Models in practical applications such as speech and image processing, bioinformatics and communication and information theory [3]. Recently there has been a renewed interest in computing the entropy rate. Entropy rate calculations based on ideas from filtering theory have been done [4], connections of entropy rate to Lyapunov exponents of random matrices have been studied in [5, 6], connections with statistical mechanics in [7, 8], and in capacity calculations of finite state channels in [9]. In this paper we follow the approach of algebraic measures [10]. Algebraic measures were introduced by Fannes, Nachtergaele and Werner in the context of quantum spin systems as classical analogues of finitely correlated states and were shown to be in one to one correspondence with functions of Markov processes or hidden Markov processes. In [11] we used the approach of algebraic measures to compute the entropy rate of a hidden Markov model with at least one unambiguous symbol and showed that an approximation to the formula converges exponentially fast to the entropy rate. Our paper is organized as follows; in section 2 we give background about the entropy rate problem, introduce the noise model and review the results of [11], in section 3 we show an efficient algorithm to compute the entropy rate and present numerical estimates of the entropy rate using a sequence of observed symbols and in section 4 we use the results to derive bounds on the capacity of the Gilbert channel.

## 2 Background

### 2.1 Setup

Consider a stationary Markov process  $\{X_1, X_2, \dots\}$  taking values in an alphabet  $N = \{0, 1, \dots, k-1\}$ . Let  $E$  be the transition matrix and  $\nu$  be the stationary Markov measure associated to the process. Let  $F_a \in M_k(\mathbb{R})$  ( $k \times k$  matrices with entries in  $\mathbb{R}$ ) be the matrix with the only non-zero row to be the  $a^{th}$  row of the transition matrix  $E$ , that is

$$(F_a)_{b,c} = \delta_{a,b} \frac{\nu((b,c))}{\nu((b))} \quad (2)$$

so that  $E = \sum_{a \in N} F_a$ . Let  $\mathbb{1} \in \mathbb{R}^k$  be the vector with all components equal to 1 and  $\tau \in \mathbb{R}^k$  be such that  $\tau_a = \nu((a))$ , the  $a^{th}$  component of the stationary distribution. The Markov measure  $\nu$  can be represented in terms of a triplet  $(\tau, \mathbb{1}, (F_a)_{a \in N}, \cdot)$ . It is easy to verify that

$$\nu((\omega_1, \dots, \omega_n)) = \langle \tau \mid F_{\omega_1} \dots F_{\omega_n} \mathbb{1} \rangle \quad (3)$$

where  $\langle u | v \rangle = u^T v$  is the usual inner product on  $\mathbb{R}^k$ . Let  $\{Y_1, Y_2, \dots\}$  with  $Y_i \in K = \{0, 1, \dots, q-1\}$  be the hidden Markov process resulting from a noisy observation of the Markov process given by the matrix  $R = [r_{ab}]$  with  $r_{ab} = Pr[Y_i = a | X_i = b]$ . One can view the output  $\{Y_n\}$  as a Markov source  $\{X_n\}$  through a discrete memoryless channel. The noisy observation of the Markov process induces a translation invariant measure  $\mu$  on  $K^{\mathbb{Z}}$  which can be written as

$$\mu(\epsilon_1, \epsilon_2, \dots, \epsilon_n) = \sum_{\substack{\omega_1, \omega_2, \dots, \omega_n \\ \omega_i \in N}} r_{\epsilon_n \omega_n} r_{\epsilon_{n-1} \omega_{n-1}} \cdots r_{\epsilon_1 \omega_1} \nu(\omega_n | \omega_{n-1}) \dots \nu(\omega_2 | \omega_1) \nu(\omega_1) \quad (4)$$

The hidden Markov process can be equivalently be represented by a function  $\Phi : N \rightarrow K$  and the measure  $\mu$  associated with can be written as

$$\mu(\epsilon_1, \epsilon_2, \dots, \epsilon_n) = \sum_{\substack{\omega_1, \omega_2, \dots, \omega_n \\ \Phi(\epsilon_i) = \omega_i}} \nu(\omega_1, \omega_2, \dots, \omega_n) \quad (5)$$

We can also represent the hidden Markov process in terms to a triplet. Let

$$E_a = \sum_{b \in L} r_{ab} F_b \quad (6)$$

It can be checked that the measure  $\mu$  can be generated by triplet  $(\tau, \mathbb{I}, (E_a)_{a \in K})$  so that

$$\mu((w_m, \dots, w_n)) = \langle \tau | E_{w_m} \dots E_{w_n} \mathbb{I} \rangle \quad (7)$$

Translation invariant measures on  $K^{\mathbb{Z}}$  which can be represented in terms of triplets were termed as manifestly positive algebraic measures in [10] and they were shown to be in one to one correspondence with functions of Markov processes or hidden Markov processes.

There is a well known formula for the entropy rate of the the Markov measure  $\nu$ . We can write the

$$H(\nu) = \sum_{a,b} \nu((a)) E_{a,b} \quad (8)$$

A tractable formula for the entropy rate of a hidden Markov process is still an open and challenging problem. Blackwell was the first to study the entropy rate of a hidden Markov process. He showed in [1] that the entropy rate of a hidden Markov process can be written as an integral of a function on a simplex with respect to a measure on the simplex. The entropy rate given by Blackwells formula is

$$H(\mu) = \sum_{a \in K} \int_{\mathcal{W}} h_a(w) \phi(dw) \quad (9)$$

and  $\phi(dw)$  is a probability measure on the simplex  $\mathcal{W} = \{(w_1, w_2, \dots, w_N) | \sum_i w_i = 1\}$  and  $h_a$  is some function on the simplex. However, practically computing the

entropy rate of a hidden Markov process using the Blackwell formula is difficult since the Blackwell measure can be hard to evaluate. Birch [2] showed that the monotonically decreasing sequence  $G_n = S(Y_n|Y_{n-1}, Y_{n-2}, \dots, Y_1)$  converges exponentially fast to the entropy rate, that is, there exist positive constants  $M$  and  $0 < \rho < 1$  such that

$$G_n - H(\mu) \leq M\rho^{n-1} \quad (10)$$

It can be seen that

$$G_n = S(Y_n, Y_{n-1}, \dots, Y_1) - S(Y_{n-1}, Y_{n-2}, \dots, Y_1) \quad (11)$$

One can compute the entropy rate using the equation 11 but it is clear that computing the entropy rate using this formula by calculating the joint probabilities involved will take time that is exponential in  $n$ .

## 2.2 Noise model and formula for the entropy rate

In [11] we considered a specific noise model which we call a hidden Markov model with at least one unambiguous received symbol. If the symbol 0 is transmitted then it is always received as 0 at the other end. On the other hand if any of the other symbol is transmitted then it is either received without any error or received as the symbol 0 with a small error probability. That is  $P(Y_i = 0|X_i = 0) = 1$ ,  $P(Y_i = 0|X_i = a) = \epsilon_a$  and  $P(Y_i = a|X_i = a) = 1 - \epsilon_a$  for  $a = 1, \dots, q-1$  and  $P(Y_i = b|X_i = a) = 0$  when  $0 \neq b \neq a$ . Here we consider the symbols  $1, 2, \dots, q-1$  to be unambiguous, since if any one of them is received then that same symbol must have been transmitted. For  $q = 2$  this model is the familiar Z-channel. See figure 1 for a description of the model in the case  $q = 2$  and  $q = 3$ . Let the matrices  $\{F_a\}$  be the matrices that describe the uncorrupted

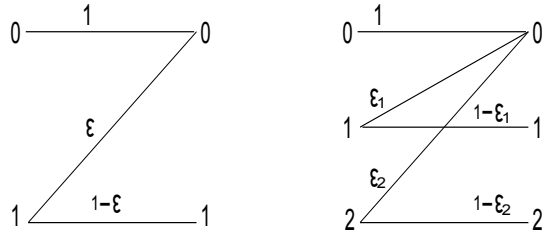


Figure 1: The noise model for  $q = 2$  and  $q = 3$ . For  $q = 2$  this noise model results in the familiar Z-channel which has been used as a model for transmission problems in optical communications. 1 and 2 are the unambiguous symbols for  $q = 3$  since if either a 1 or a 2 was received then we can conclude with certainty that the sent symbol was the same. If a 0 is received then all of the three symbols could have been transmitted; 1 and 2 with probability  $\epsilon_1$  and  $\epsilon_2$  and 0 with probability  $1 - \epsilon_1 - \epsilon_2$ .

Markov source as in equation (3). For this noise model we write the matrices  $\{E_a\}$  given by equation (6) as

$$\begin{aligned} E_0 &= F_0 + \sum_{a=1}^{q-1} \epsilon_a F_a \\ E_a &= (1 - \epsilon_a) F_a \quad \text{for } a = 1, \dots, q-1 \\ \sum_{a \in K} E_a &= E \end{aligned} \tag{12}$$

Let  $\Gamma_a : \mathcal{W} \rightarrow \mathcal{W}$  be a mapping on the simplex  $\mathcal{W}$  defined by

$$\Gamma_a(\nu) = \frac{E_a^T \nu}{\langle \nu E_a \mid \mathbb{1} \rangle} \tag{13}$$

Let  $e_i, i = 0, 1, \dots, q-1$  denote the transpose of the  $(i+1)^{st}$  row of  $E = [e_{ij}]$ . In [11] we showed that the support of the Blackwell measure for the hidden Markov model described by the noise model in this section is countable.

**Proposition 2.1** ([11]). *For the HMP with one or more unambiguous received symbol the support of the measure  $\phi$  is given by*

$$\Delta = \overline{\{\Gamma_0^m e_j \mid j \in \{1, \dots, q-1\}; m \in \mathbb{N}_0\}} \tag{14}$$

Next we state the assumptions and statement of the main theorem from [11] for the entropy rate of the hidden Markov process under consideration. Let  $p = \min_{ij} e_{ij}$  and  $P = \max_{ij} e_{ij}$ .

*Assumption 1 :*

- i)  $0 < p \leq P < 1, \epsilon_0 = 1, \epsilon_a > 0 \quad \forall a \in \{1, \dots, q-1\}$
- ii)  $E_0$  is a one to one mapping

Define

$$c_{j,m} = \prod_{i=1}^m \langle \Gamma_0^{m-i} e_j \mid E_0 \mathbb{1} \rangle \tag{15}$$

Let  $A$  be the  $q \times q-1$  matrix defined by entries.

$$\begin{aligned} A_{ij} &= -\delta_{ij} + \sum_{m=0}^{\infty} \langle \Gamma_0^m e_j \mid E_i \mathbb{1} \rangle c_{j,m} \quad \text{if } i \neq q, q \neq 2 \\ A_{ij} &= 0 \quad \text{if } i \neq q, q = 2 \\ A_{qj} &= \sum_{m=0}^{\infty} c_{j,m} \end{aligned} \tag{16}$$

$$\Phi = [\phi(e_1) \cdots \phi(e_{q-1})]^T \in \mathbb{R}^{q-1}, b = [0 \ 0 \cdots 1]^T \in \mathbb{R}^q$$

Here  $\phi(e_i)$  is the weight of measure  $\phi$  at the point  $e_i \in \mathbb{R}^q$ . Let  $h_a : \mathcal{W} \rightarrow \mathbb{R}$  be the function defined as

$$h_a(\nu) = -\langle \nu \mid E_a \mathbb{1} \rangle \log \langle \nu \mid E_a \mathbb{1} \rangle \quad (17)$$

**Theorem 2.2** ([11]). *Under Assumption 1 the entropy rate of the measure  $\mu$  associated with the hidden Markov process with the noise model described in this section is given by*

$$H(\mu) = \sum_{j=1}^{q-1} \sum_{m=0}^{\infty} \sum_{a=0}^{q-1} h_a(\Gamma_0^m e_j) c_{j,m} \Phi_j \quad (18)$$

In [11] we showed that an approximation to the formula for  $H(\mu)$  converges exponentially fast to the entropy rate. For the HMP under consideration the result for the exponential convergence was much more simpler to show than Birch's general result given by equation 10. For the approximation to  $H(\mu)$  let

$$A = \hat{A} + R \quad (19)$$

where the entries of  $R$  are the tails  $((N+1)^{st}$  term onwards) of the entries of  $A$ . Let  $\hat{\Phi}$  be the least square solution to

$$\hat{A}\hat{\Phi} = b \quad (20)$$

$$\text{Therefore } \hat{\Phi} = \hat{A}^\dagger b \quad (21)$$

where  $\hat{A}^\dagger = (A^T A)^{-1} A^T$  is the pseudo-inverse of  $A$ . Define

$$H_N(\mu) = \sum_{j=1}^{q-1} \sum_{m=0}^N \sum_{a=0}^{q-1} h_a(\Gamma_0^m e_j) c_{j,m} \hat{\Phi}_j$$

$$\text{and } \gamma := \max_j \sup_k \sum_{a=0}^{q-1} \epsilon_a [\Gamma_0^k e_j]_a \quad (22)$$

We have the following theorem

**Theorem 2.3** ([11]). *Under Assumption 1 the entropy rate  $H(\mu)$  of the hidden Markov process with the noise described can be approximated to  $O(\gamma^{N+1})$  by  $H_N(\mu)$  and we have*

$$|H(\mu) - H_N(\mu)| \leq B \gamma^{N+1} \quad \text{with } B = \frac{q}{1-\gamma} \left(1 + \frac{q \|\hat{A}^\dagger\|_1}{1-\gamma}\right) \quad (23)$$

### 3 Estimation and computation of entropy rate

#### 3.1 An efficient algorithm to compute entropy rate

Birch's result (equation 10) shows that the computation of the entropy rate of a general hidden Markov chain using the monotonically decreasing sequence  $G_n$  converges exponentially to the actual entropy rate. However the computation of  $G_n$  using equation 11 method takes exponential time in  $n$ . In this section we give a  $O(Nq^3)$  algorithm to compute the entropy rate of the hidden Markov model with unambiguous symbols using the approximate formula  $H_N(\mu)$ . If  $\delta = |H(\mu) - H_N(\mu)|$  is accuracy with which we compute  $H(\mu)$  then we get an algorithm that is  $O(\log \frac{1}{\delta})$  in terms of the accuracy as compared to  $O(\frac{1}{\delta})$  if we use the brute force formula  $G_n$ . In this section we prove these results and substantiate them with numerical computations. The algorithm to compute the entropy rate is as follows:

**Algorithm I:**

**Inputs:**

- i) A  $q \times q$  transition matrix  $E$  of the Markov chain  $E_{ij} = P(X_n = j | X_{n-1} = i)$ .
- ii) A  $q \times q$  channel probability matrix  $R$  with  $R_{ab} = P(Y_n = a | X_n = b)$  according to the hidden Markov process under consideration.
- iii)  $N$  the number of terms of the approximate formula.

Both  $E$  and  $R$  should satisfy conditions specified by Assumption 1.

**Output:** The entropy rate  $H$  of the hidden Markov model.

**Step 1:**

From the matrices  $E$  and  $R$  construct matrices  $E_0$  according to equation 12.

For  $j = [0 \cdots q - 1]$  and  $m = [1 \cdots N]$  compute  $\Gamma_0^m e_j$ . (where  $\Gamma_0$  is given by equation 13 and  $e_j$  is the transpose of the  $(i + 1)^{st}$  row of  $E$ )

**Step 2:**

For  $j = [0 \cdots q - 1]$  and  $m = [1 \cdots N]$  compute  $c_{j,m} = \Pi_{i=1}^m \langle \Gamma_0^{m-i} e_j | E_0 \mathbb{I} \rangle$ .

**Step 3:**

Compute the entries of the matrix  $\hat{A}$  given by equation 19, the pseudo-inverse  $\hat{A}^\dagger = (A^T A)^{-1} A^T$  and then vector  $\hat{\Phi} = \hat{A}^\dagger b$ .

**Step 4:**

Using the precalculated values of  $\Gamma_0^m e_j$ ,  $c_{j,m}$  and  $\Phi_j$  in Steps 1,2 and 3 do the following computation.

$H = 0$ .

For  $j \in [1 \cdots q - 1]$ ,  $m \in [0 \cdots N]$  and  $a \in [0 \cdots q - 1]$

$H = H + -(\Gamma_0^m e_j)_a \log(\Gamma_0^m e_j)_a c_{j,m} \hat{\Phi}_j$ .

Output  $H = H_N(\mu)$  as the entropy rate.

**Theorem 3.1.** *Run time complexity of Algorithm I to compute the entropy rate is  $O(Nq^3)$*

*Proof.* We analyze the steps of Algorithm I

- Each computation in *Step 1* is matrix multiplication of a  $q \times q$  matrix  $\Gamma_0$  with a  $q \times 1$  vector  $e_j$  which using a standard matrix multiplication

algorithm requires  $O(q^2)$  time. There are  $Nq$  total such computations and hence the time complexity of *Step 1* is  $O(Nq^3)$ .

- In *Step 2* one requires the computation of  $c_{j,m} = \Pi_{i=1}^m \langle \Gamma_0^{m-i} e_j \mid E_0 \mathbb{I} \rangle$ .  $E_0 \mathbb{I}$  takes  $O(q)$  time and for each  $j$  the inner product  $\langle \Gamma_0^k e_j \mid E_0 \mathbb{I} \rangle$  can be done in  $O(q)$  time.  $c_{j,k}$  can be computed iteratively as  $c_{j,k} = c_{j,k-1} \langle \Gamma_0^k \mid E_0 \mathbb{I} \rangle$  and since there are  $Nq$  such computations total time taken by *Step 2* is  $O(Nq^2)$ .
- In *Step 3* we first compute the matrix elements of the  $q \times q - 1$  matrix  $\hat{A}$ . Each term of  $\hat{A}$  is given by equation 19 up till the first  $N$  terms. Each matrix entry thus requires  $O(Nq)$  time and since there are order  $q^2$  terms computing  $\hat{A}$  requires  $O(Nq^3)$  time. Next we compute the pseudo inverse  $\hat{A}^\dagger = (A^T A)^{-1} A^T$  which is a combination of matrix multiplication and taking inverse which by standard methods takes  $O(q^3)$  time. Computing  $\Phi = \hat{A}^\dagger b$  requires  $O(q^2)$  time, hence the total time required in *Step 3* is  $O(Nq^3)$ .
- Finally *Step 4* has  $Nq^2$  basic operations of addition or multiplication and hence requires  $O(Nq^2)$  time.

From the above analysis we get that the time complexity of Algorithm I is  $O(Nq^3)$ .  $\square$

**Theorem 3.2.** *The running time of Algorithm I to compute  $H_N(\mu)$  to within  $\delta$  accuracy of  $H(\mu)$  is  $O(\log \frac{1}{\delta})$*

*Proof.* From the bound of equation 2.3 we get that

$$|H(\mu) - H_N(\mu)| \leq B\gamma^{N+1} \quad \text{with } B = \frac{q}{1-\gamma} \left(1 + \frac{q\|\hat{A}^\dagger\|_1}{1-\gamma}\right)$$

Therefore to obtain a  $\delta$  accuracy in computation of  $H(\mu)$  we need

$$\begin{aligned} \delta &\leq B\gamma^{N+1} && \text{that is} \\ \frac{1}{\delta} &\geq B\gamma^{N+1} \\ \log\left(\frac{1}{\delta}\right) &\geq \log B + (N+1)\log \gamma \end{aligned}$$

dividing by the negative quantity  $\log \gamma$  gives

$$N+1 \geq \frac{\log\left(\frac{1}{\delta B}\right)}{\log \gamma} \tag{24}$$

Combining with theorem 3.1 we get that the time complexity of Algorithm I to compute  $H_N(\mu)$  to  $\delta$  accuracy is  $O(\log \frac{1}{\delta})$ .  $\square$



We present a numerical example for approximating the entropy rate formulas given by theorem 2.3. Let  $q = 3$ ,  $\epsilon_1 = 0.01$  and  $\epsilon_2 = 0.02$ . The transition matrix we use is

$$E = \begin{pmatrix} 0.4 & 0.25 & 0.35 \\ 0.25 & 0.45 & 0.3 \\ 0.2 & 0.55 & 0.25 \end{pmatrix}$$

The results of the entropy rate calculations are seen in table 1. A comparison with calculations of the entropy rate for the same HMP but by using the brute force formula of equation 1 is seen in table 2.

N	$H_N(\mu)$	err(N) bound	Execution time (in secs)
10	1.520946691296695	0.3561	0.0077
20	1.520947864830033	0.0030	0.0129
30	1.520947864969799	$2.6758 \times 10^{-5}$	0.0197
40	1.520947864969815	$2.3193 \times 10^{-7}$	0.0278
50	1.520947864969815	$2.0103 \times 10^{-9}$	0.0289

Table 1: The estimated entropy rate  $H_N(\mu)$  using the formula given by theorem 2.3.

n	$S_n(\mu) - S_{n-1}(\mu)$	Execution time (in secs)
5	1.520946036478195	0.0581
6	1.520947599473784	0.127
7	1.520947829277763	0.342
8	1.520947860073111	1.08
9	1.520947864301537	3.479
10	1.520947864877943	11.14

Table 2: The estimated entropy rate using the brute force formula  $H(\mu) = S_n(\mu) - S_{n-1}(\mu)$ .

### 3.2 Estimating the entropy rate from an observed sequence

In the previous subsection we have assumed that the transition matrix  $E$  of the Markov chain and the noise parameters  $\epsilon_a$  are known. However in many practical applications this is not the case. In this section we assume that we are only given an observation sequence and we have to estimate the entropy rate. In this method we use parameter estimation to estimate the transition matrix and the noise parameters  $\epsilon_a$  and then use Algorithm I to compute the entropy rate. Let  $Y$  be a vector of the observed symbols for time  $t = 1$  to  $t = N$  and let  $X$  be the corresponding hidden or state symbols and let  $Z = \begin{pmatrix} Y \\ X \end{pmatrix}$ . Let the

unknown parameters be represented by  $\theta = \begin{pmatrix} \eta \\ \epsilon \end{pmatrix}$  where  $\eta$  is a vector containing the unknown transition matrix entries and  $\epsilon$  is the vector representing the noise parameters. We have

$$p(Z|\theta) = p(Y, X|\theta) = p(Y|X, \epsilon)p(X|\eta)$$

Assuming the initial distribution of the Markov chain is uniform we get

$$p(Z|\theta) = \frac{1}{q} \prod_{t=1}^N p(Y(t)|X(t), \epsilon) \eta_{t,t+1}$$

and the log-likelihood function  $L(Z|\theta)$  and the complete likelihood function  $Q(\theta|\theta')$  respectively

$$\begin{aligned} L(Z|\theta) &= \log p(Z|\theta) = \sum_{t=1}^N p(Y(t)|X(t), \epsilon) + \log \eta_{t,t+1} - \log q \\ Q(\theta|\theta') &= \sum_{X \in K^n} L(Z|\theta) p(X|Y, \theta') \end{aligned} \quad (25)$$

To compute the complete likelihood function the  $Q(\theta|\theta')$  conditional probabilities  $p(X|Y, \theta')$  need to be estimated. This can be done using the Baum-Welsh forward-backward algorithm [3]. Since the Markov input sequence it is only required to estimate the probabilities

$$p(X(t) = k, X(t+1) = l|Y) = \frac{p(X(t) = k, X(t+1) = l, Y)}{P(Y)}$$

For each  $t$  we define row vector  $1 \times q$  vector  $\alpha(t)$ ,  $q \times 1$  vector  $\beta(t)$  and  $q \times q$  matrix  $m(t)$

$$\begin{aligned} \alpha_k(t) &= p(X(t) = k, Y_1^{t-1} = y_1^{t-1}) \\ \beta_t(l+1) &= p(Y_{t+1}^N | X(t+1) = l) \\ m_{kl} &= p(Y(t)|X(t) = k)P(X(t+1) = l|X(t) = k) \end{aligned}$$

where we use the notation  $p(Y_1^N) = p(Y_1, \dots, Y_N)$ , then one can see that

$$P(X(t) = k, X(t+1) = l, Y) = \alpha_k m_{kl}(t) \beta_l(t+1) \quad (26)$$

and we observe the following forward and backward recursion equations

$$\begin{aligned} \alpha(t+1) &= \alpha(t)m(t) \\ \beta(t) &= m(t)\beta(t+1) \end{aligned} \quad (27)$$

if  $\gamma(t) = \alpha(t)\beta(t)$  then we have

$$\gamma(t) = \alpha(t)m(t)\beta(t+1) = \alpha(t+1)\beta(t+1) = \gamma(t+1)$$

that is  $\gamma(t)$  is time invariant and in fact

$$p(Y = y) = \sum_{k \in K} \sum_{l \in K} p(X(t) = k, X(t+1) = l, Y = y)$$

So that from equations 26 and recursion equations 27 we get

$$\begin{aligned} p(Y = y) &= \sum_{k \in K} \sum_{l \in K} p(X(t) = k, X(t+1) = l, Y = y) \\ &= \sum_k \alpha_k(t) \beta_k(t) \end{aligned}$$

therefore

$$p(Y = y) = \alpha(t) \beta(t) = \gamma(t) \quad (28)$$

Equations 26 and 28 can be used to estimate  $p(X|Y, \theta')$ . One can start with a guess of  $\alpha(0) = [\frac{1}{q} \dots \frac{1}{q}]$  and  $\beta(N+1) = [1, \dots, 1]'$  and then iterate using 27 to get the values of  $\alpha(1), \dots, \alpha(N)$  and  $\beta(N), \dots, \beta(1)$ . We can substitute equations 26 and 28 to see that

$$p(X|Y, \theta') = \frac{\alpha_k(t, \theta') m_{kl}(t, \theta') \beta_l(t, \theta')}{\gamma(\theta')} \quad (29)$$

The expectation maximization algorithm involves two steps. After making an initial guess of parameter  $\theta' = \theta_0$  and setting maximum number of iterations  $k$  and a tolerance level for the successive estimates  $\delta$  we have

i Expectation Step

Use the Baum-Welsh forward backward algorithm described above to compute the conditional probabilities  $p(X|Y, \theta')$  and complete likelihood function  $Q(\theta|\theta_j)$ .

ii Maximization Step

Set the new value of  $\theta'$

$$\theta_{j+1} = \max_{\theta} Q(\theta|\theta_j)$$

The maximization can be done analytically using Lagrange multipliers or computed numerically. If  $\|\theta_{j+1} - \theta_j\| > \delta$  and number of iterations are less  $k$  than go to step 1 otherwise set  $\theta = \theta_{j+1}$ .

We generated 200 output symbols using  $E = \begin{pmatrix} 0.25 & 0.35 & 0.4 \\ 0.15 & 0.45 & 0.4 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$  and  $\epsilon_1 = 0.02$

and  $\epsilon_2 = 0.03$ . The EM algorithm gives  $\hat{E} = \begin{pmatrix} 0.224 & 0.323 & 0.453 \\ 0.113 & 0.476 & 0.411 \\ 0.24 & 0.299 & 0.46 \end{pmatrix}$  and  $\hat{\epsilon}_1 = 0.048$  and  $\hat{\epsilon}_2 = 0.042$ . The entropy estimate using Algorithm I with  $N = 100$ ,  $\hat{E}$ ,  $\hat{\epsilon}_1$  and  $\hat{\epsilon}_2$  is 1.51808 and which is close to the estimate of 1.51715 using transition matrix E and noise parameters  $\epsilon_1$  and  $\epsilon_2$ .

## 4 Bounds on the capacity of the Gilbert channel

The Gilbert channel [12] is a channel with memory which is used to model burst errors. The channel state  $S_n$  at time  $n$  can be good (G) or bad (B) and the channel transitions between good and bad states according to a Markov chain. When the channel is in a good state the input bit is transmitted without error and when the channel is in a bad state there is a probability of a bit flip is  $h$ . To model burst errors the channel is modeled so that transition probability from a good state to a bad state ( $P$ ) and bad state to good state are small ( $Q$ ). The output at time  $n$  is given by  $Y_n = X_n + Z_n$  where  $X_n$  is the input at time  $n$  and  $Z_n$  is the noise and the addition is modulo 2. The noise  $Z_n$  will be 0 if the state of the channel is in good state and if the channel is in B state then the noise will be 0 or 1 decided on a coin flip with bias  $h$ . The noise process  $Z_n$  can be looked at as the output of the Z-channel with Markov input  $S_n$  and with transition matrix  $E = \begin{pmatrix} 1-P & P \\ Q & 1-Q \end{pmatrix}$  (see figure 2). The capacity of such a

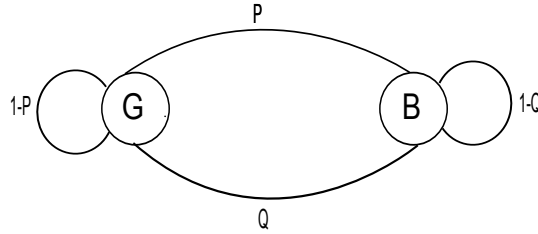


Figure 2: The Gilbert channel is a channel with memory. The channel state at time  $n$   $S_n$  transitions like a Markov chain between a good state (G) and a bad state (B). The transition probabilities from G state to B state is  $P$  and from B to G is  $Q$ . In the good state the channel acts like a perfect channel while in the bad state there is a probability of a bit flip  $h$ . The output of the Gilbert channel can be written as  $Y_n = X_n \oplus Z_n$  where  $X_n$  is input process and  $Z_n$  is output of the Z-Channel with the channel state Markov process  $S_n$  as input.

finite state channel is defined as

$$C = \lim_{n \rightarrow \infty} \frac{1}{n} \max_{p(X^n)} I(X^n; Y^n)$$

where  $I(X^n; Y^n)$  is the mutual information between the the input and output sequences. The capacity of finite state Markov channels have been studied in [13, 14]. We use the results of the previous section to obtain bounds on the capacity of the Gilbert channel. Let  $H_N$  be the approximate formula for the entropy rate of the Z-Channel given by theorem 2.3 with  $q = 2$  and  $B$  be defined as in equation 23 then we have the following theorem:

**Theorem 4.1.** *Under Assumption 1 the capacity  $C$  of the Gilbert channel with*

channel noise process  $\{Z_n\}$  can be upper and lower bounded for all  $N \in \mathbb{N}$  as

$$1 + H_N(\mu) - B\gamma^{N+1} \leq C \leq 1 + H_N(\mu) + B\gamma^{N+1}$$

*Proof.* It can be shown that the capacity of the Gilbert channel is

$$C = 1 - H(Z)$$

where  $H(Z)$  is the entropy rate of the noise process  $\{Z_n\}$ . Indeed we can write

$$I(X^n; Y^n) = \sum_{i=1}^N H(Y_i | Y_1^{i-1}) - H(Y_i | X_i, X_1^{i-1}, Y_1^{i-1})$$

Due to the relation  $Y_n = X_n \oplus Z_n$  between the input, output and noise we observe that

$$H(Y_i | X_i, X_1^{i-1}, Y_1^{i-1}) = H(Z_i | Z_1^{i-1})$$

Since Markov channel state process  $S_n$  is independent of the input, and the noise process  $Z_n$  is the hidden Markov process;  $Z_n = \phi(S_n)$  for some function  $\phi$  therefore  $H(Z_i | Z_1^{i-1})$  is independent of the input distribution  $p(X^n)$ . Thus

$$\begin{aligned} C &= \lim_{n \rightarrow \infty} \frac{1}{n} \max_{p(X^n)} \left( \sum_{i=1}^n H(Y_i | Y_1^{i-1}) - H(Y_i | X_i, X_1^{i-1}, Y_1^{i-1}) \right) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \max_{p(X^n)} \left( \sum_{i=1}^n H(Y_i | Y_1^{i-1}) - H(Z_i | Z_1^{i-1}) \right) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \max_{p(X^n)} \sum_{i=1}^n H(Y_i | Y_1^{i-1}) - \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(Z_i | Z_1^{i-1}) \end{aligned}$$

Therefore

$$C = \lim_{n \rightarrow \infty} \frac{1}{n} \max_{p(X^n)} \sum_{i=1}^n H(Y_i | Y_1^{i-1}) - H(Z) \quad (30)$$

Now consider

$$\begin{aligned} H(Y_i | Y_1^{i-1}) &= \\ - \sum_{y_1, \dots, y_{i-1}} \left( \sum_y p(Y_i = y | y_1, \dots, y_{i-1}) \log p(Y_i = y | y_1, \dots, y_{i-1}) \right) p(y_1, \dots, y_{i-1}) \end{aligned} \quad (31)$$

The i.i.d uniform input distribution maximizes  $p(Y_i = y | y_1, \dots, y_{i-1})$ . Indeed

$$p(Y_i = y | y_1, \dots, y_{i-1}) = \sum_{s_i} p(Y_i = y | S_i = s_i) p(S_i = s_i | y_1, \dots, y_{i-1}) \quad (32)$$

Also,

$$p(Y_i = y | S_i = s_i) = \sum_x p(Y_i = y | X_i = x, S_i = s_i) p(X_i = x)$$

Due to the symmetry of the channel  $\sum_x p(Y_i = y|X_i = x, S_i = s_i)$  is independent of  $Y_i$  therefore for the uniform i.i.d. input distribution the conditional density  $p(Y_i = y|S_i = s_i)$  is a constant and is equal to  $\frac{1}{2}$ . Substituting this in 32 we get that  $p(Y_i = y|y_1, \dots, y_{i-1}) = \frac{1}{2}$ . The quantity

$$\sum_y p(Y_i = y|y_1, \dots, y_{i-1}) \log p(Y_i = y|y_1, \dots, y_{i-1})$$

in equation 31 for the uniform i.i.d. input distribution gets maximized to  $2\frac{1}{2} \log 2 = 1$ . Therefore from equation 30 we get

$$C = 1 - H(Z)$$

Using theorem 2.3 we can bound the capacity of the Gilbert channel

$$1 + H_N(\mu) - B\gamma^{N+1} \leq C \leq 1 + H_N(\mu) + B\gamma^{N+1}$$

□

We assume the transition matrix  $E = \begin{pmatrix} 0.8 & 0.2 \\ 0.25 & 0.75 \end{pmatrix}$  for the channel transitions compute the capacity of the Gilbert channel using different values of  $h$  parameter.

h	C (lower bound)	C (upper bound)
0.02	1.775537282409934	1.775537393396272
0.04	1.787283765040533	1.787284687386383
0.06	1.797179493884635	1.797187010251415
0.08	1.805422838236253	1.805483229491273
0.1	1.812015925779448	1.812497634488852

Table 3: The upper and lower bounds on the capacity of the Gilbert Channel computed using the entropy rate formula

## References

- [1] D. Blackwell, "The entropy of functions of finite-state Markov chains," *Trans. 1st Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, pp. 13–20, 1957.
- [2] J. Birch, "Approximations for the entropy for functions of Markov chains," *Ann. Math. Statistics*, vol. 33, pp. 930–938, 1962.
- [3] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of IEEE*, vol. 77, pp. 257–286, 1989.

- [4] E. Ordentlich and T. Weissman, “On the optimality of symbol by symbol filtering and denoising,” *IEEE Trans. Inf. Theory*, vol. 52, pp. 19–40, 2006.
- [5] P. Jacquet, G. Seroussi, and W. Szpankowski, “On the entropy of a hidden Markov process,” *Proceedings of Data Compression Conference, Snowbird, UT*, pp. 362–371, 2004.
- [6] T. Holliday, A. Goldsmith, and P. Glynn, “On entropy and lyapunov exponents for finite state channels,” *IEEE Trans. Inf. Theory*, vol. 52, 2006.
- [7] O. Zuk, E. Domany, I. Kanter, and M. Aizenmann, “From finite-system entropy to entropy rate of a hidden Markov process,” *IEEE Signal Processing Letters*, vol. 13, pp. 517–520, 2006.
- [8] Allahverdyan, “Entropy of a hidden Markov process via cycle expansion,” *Journal of Statistical Physics*, vol. 133, pp. 535–564, 2008.
- [9] H. Pfister, “On the capacity of finite state channels and the analysis of convolutional accumulate-m codes,” *PhD thesis, University of California, San Deigo*, 2003.
- [10] M. Fannes, B. Nachtergaele, and L. Slegers, “Functions of Markov processes and algebraic measure,” *Reviews in Mathematical Physics*, vol. 4, p. 39, 1992.
- [11] K. Marchand, J. Mulherkar, and B. Nachtergaele, “Entropy rate calculations using algebraic measures,” *IEEE Int. Sym. on Inf. theory proceedings, Boston USA*, pp. 1072–1076, 2012.
- [12] E. Gilbert, “Capacity of burst-noise channel,” *Bell System Technical journal*, vol. 39, pp. 1253–1265, 1960.
- [13] A. Goldsmith and P. Varaiya, “Capacity, mutual information and coding for finite state Markov channels,” *IEEE Trans. Inf. Theory*, vol. 42, pp. 2779–2784, 1996.
- [14] M. Rezaeian, “Symmetric characterization of finite state Markov channels,” *IEEE Int. Sym. on Inf. theory proceedings, Seattle USA*, pp. 2734–2738, 2006.